1998

The FGDC Feature Registry and its Role in Supporting Semantic Interoperability

Kevin Backe and Demetra Voyadgis

U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road, Alexandria, VA 22315-3864

kevin.backe@usace.army.mil
dvoyadgi@tec.army.mil

Abstract. This paper discusses the Federal Geographic Data Committee's (FGDC) Feature Registry project and what the FGDC is doing to support semantic interoperability. It also discusses the semantic information captured for geospatial data by U.S. Federal Government agencies and their current practices for semantic mapping. This paper explores current, ongoing research and development that may contribute to enhancing the semantic interoperability for the geospatial data created by the U.S. Federal Government. Finally, this paper calls for additional research on semantic mapping that hopefully will lead to simplifying what has been appropriately labeled a "hard problem."

Introduction

The increased use of geospatial data creates an increased need to share these data. Geospatial data creation is costly, compelling users to acquire existing data. The merging of geospatial data from disparate sources often has unforeseen difficulties. Prime among these is semantic interoperability. This paper will discuss the problem of semantic interoperability. It will explain the role of the Federal Geographic Data Committee (FGDC) in this area, and the FGDC's Feature Registry Project. It will examine the current research being performed, and recommend additional research needed to obtain semantic interoperability.

Semantic Interoperability

Semantic interoperability, in particular the irreconcilable difference of feature's and attribute's appellations between data sets, manifests three types of problems for geospatial users. The first is the general inability to share data. The second is the development of software applications that are specific to certain semantics. The third problem is non-extensible queries.

Semantic heterogeneity is one of the primary limitations that hamper the widespread ability to share geospatial data. Much of the existing geospatial data contains semantic information that is unique to each database. This is especially true of organizations that generate geospatial data for state-, county-, or municipality-level applications, as well as for private industries--such as utility companies. At the Federal level, many of the agencies that generate geospatial data have "agency specific" semantics. These Federal agencies may use the same semantics for a series of geospatial databases (or products). However, these semantics are too narrowly focused to be used by other data producers. As a result, these "agency specific" semantics are only used for a limited number of databases. In addition, these "agency specific" semantics are often not published, and are not readily available to the consumers of these products, especially the public.

Software applications that are developed to exploit geospatial data are tailored to these "database specific" or "agency specific" semantics, and are limited to a number of data sets. For example, an application that
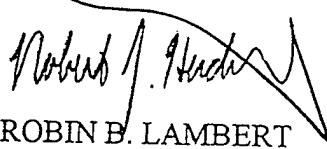
19981020 054

16 September 1998

MEMORANDUM FOR Kevin Backe and Demetra Voyadgis

SUBJECT: Clearance of Conference Paper

The following conference paper, "The Federal Geographic Data Committee's (FGDC) Feature Registry and its Role in Supporting Semantic Interoperability," written by Kevin Backe and Demetra Voyadgis, TEC, to be presented at the 2nd International Conference on Interoperating Geographic Information Systems, 10-12 March 1999, in Zurich, Switzerland, has been reviewed by this office, and by TEC's Security Office, and pending approval from HQ, USACE, is cleared for presentation.

Encl

for ROBIN B. LAMBERT
Chief, Technical Plans
and Programs Office

M. BISHOP
SO

assesses vehicular mobility across the open terrain is tailored to available feature and attribute types, and domain values captured in slope, soils, and vegetation information. To use these applications on data with different semantics, it is necessary either to make changes to these applications, or translate the semantics contained in this new geospatial data source into the application's semantics. This is often a very cumbersome (perhaps impossible) process; consequently, new geospatial data sources are rarely used.

In addition, queries that rely on semantic information must be tailored to the semantic information contained in a given data set. For example, a query on a database for all "primary, divided highways in Omaha, NE" must query on those exact terms; therefore, in order to generate a query that uses semantic information contained in a geospatial database, it is necessary either to know, or expose and examine, the semantics of the database.

Role of the FGDC in Semantic Interoperability

The FGDC was established to support the National Spatial Data Infrastructure (NSDI), which includes providing the necessary geospatial data standards to help enable the interoperability of geoprocessing systems and geospatial data sharing. The FGDC established 18+ thematic subcommittees and working groups with the primary focus of developing geospatial data standards for their specified information communities. The following is a list of these FGDC subcommittees and working groups:

> Subcommittee for Base Cartographic Data, Bathymetric Subcommittee, Cadastral Subcommittee, Subcommittee on Cultural and Demographic Data, Federal Geodetic Control Subcommittee, Geologic Subcommittee, Ground Transportation Subcommittee, Subcommittee on International Boundaries & Sovereignty, Soils Subcommittee, Vegetation Subcommittee, Water Subcommittee, Wetlands Subcommittee, Biological Data Working Group, Earth Cover Working Group, Facilities Working Group, Historical Data Working Group, Metadata Ad Hoc Working Group, Sample Inventory and Monitoring of Natural Resources and the Environment Working Group.

To date there are 30+ standards being developed by these FGDC subcommittees and working groups. The majority of these FGDC standards address various communities' semantics (i.e., metadata profiles, data content, and classification standards). The development and adoption of these standards by Federal agencies and other FGDC partner organizations that produce geospatial data (i.e., state agencies, local and tribal governments, ...) will make the semantic interoperability problem a more manageable problem.

With the development and use of "community-specific" semantic geospatial data standards, it becomes possible to share geospatial information within a community. The semantics contained within these data have a shared meaning within that community. However, there are large amounts of legacy geospatial data, at best, conforming to agency's or community's semantics. In addition, multiple standards are being developed simultaneously by different groups/communities. All these standards are at different stages in the development process (and employ many different data models). Consequently, it has been very difficult for the FGDC to coordinate the content of all these standards; therefore, the FGDC Standards Working Group established a Feature Registry with the intent of examining the relationship between geospatial data content and classification standards.

The FGDC Feature Registry Project

This Feature Registry project is being co-sponsored by the U.S. Army Corps of Engineers and the FGDC Standards Working Group and is being coordinated with the FGDC thematic subcommittees and working groups that are in the process of developing (or have developed) thematic data content and classification standards. Specifically, the FGDC Feature Registry is a repository for feature/attribute/domain information available from FGDC data content and classification standards.

The primary purpose of the FGDC Feature Registry is to serve as a single repository for geospatial data content and classification standards that will allow the FGDC to easily identify potential overlaps and conflicts in the data content standards currently being developed independently by FGDC subcommittees, working groups, and agencies. In addition, a national feature registry that integrates multiple thematic disciplines would support a broad base of applications that require cross-theme geographic analysis, and will enhance data sharing opportunities across Federal and non-federal user communities.

The short-term objective of the FGDC Feature Registry project is to populate the thematic feature registry with content and classification standards. The long-term objective is to build a thesaurus that will allow cross-thematic links, and resolve conflicts, where possible, across themes.

To date, the following FGDC standards have been incorporated into the Registry: the FGDC Vegetation, Soils, and Wetlands Classification Standards, and the Utilities, Environmental Hazards, and Hydrographic Data Content Standards. The FGDC also has actively initiated incorporation of other significant geospatial communities' feature dictionaries (catalogs) into the Feature Registry, especially those dictionaries that have been used to capture the semantics for a significant volume of geospatial data, for example:

> the North Atlantic Treaty Organization's (NATO) Feature Attribute Coding Catalog; the International Hydrographic Organization's (IHO) Digital Hydrographic Data Object Catalog; the U.S. Geological Survey's (USGS) Digital Line Graphic–Enhanced dictionary for topographic products; the American National Standards Institute's (ANSI) SDTS part 2; the U.S. Census Bureau's Tiger line dictionary; and the U.S. Department of Defense's Tri-Service Spatial Data Standard.

Part of the task of incorporating additional standards into the Feature Registry requires transforming them into the registry data model. Also, generating an integration report is required to verify the results of this integration (including documenting any open issues with the integration of each standard into the Feature Registry database).

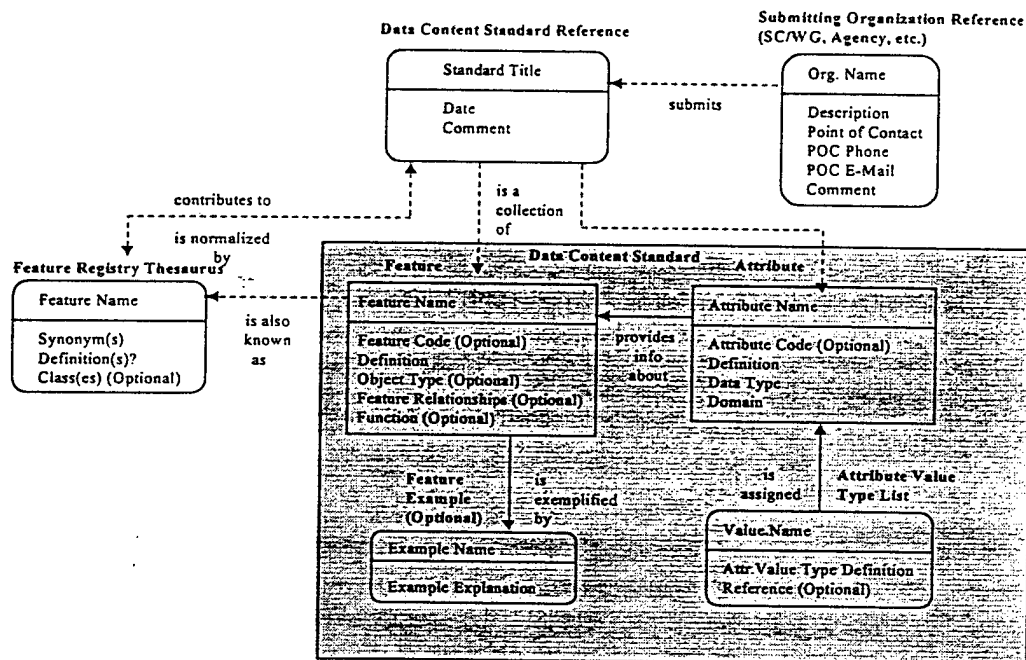## FGDC FEATURE REGISTRY MODEL (THEMATIC VIEW)

Fig. 1. FGDC Feature Registry Logical Data Model

In addition to its primary purpose, the FGDC Feature Registry is well positioned to capture the semantic mapping between available feature dictionaries. The Feature Registry is developing a "thesaurus" concept to associate (and link) related feature/attribute information contained within the Registry, which has been furnished from various standards/dictionaries. With proper enhancements, this thesaurus approach to mapping between related feature/attribute information could be valuable in supporting geospatial data translation tools.

Several software tools have been developed for the Feature Registry project. The Feature Registry itself is based on an MS Access database technology using a common feature/attribute/domain geospatial data model (which is the logical data model defined in the Spatial Data Transfer Standard, and in the Draft ISO TC211 Standard 15046-10 for Geographic Information/Geomatics Feature Cataloging Methodology). A Feature Registry "Loader" tool has been developed and is available to input a geospatial data content or classification standard that can be directly incorporated into the registry. Also, a Feature Registry "Query" tool has been developed to provide a capability to examine the registry and to aid in the process of discovering potentially related terms contained within the Feature Registry. This Query tool has been developed as both a web-based application that can access a web-based version of the registry, as well as a downloadable application that can query a downloadable version of the Feature Registry database.

Thankfully, some previous mapping work has been done for several of these standards. However, in general, the tedious task of mapping information between geospatial community semantics is rarely performed. And, if a mapping was done, this information is rarely published and often not well documented and maintained, especially with new/updated versions of each standard. Moreover, the software tools available today to identify the potentially related terms for semantic mapping are very limited. Currently, the Feature Registry project employs a simple word match capability and other manual methods to perform this semantic mapping between feature catalogs. Obviously this process is very laborious, wherein lies the need for more sophisticated tools to enhance the automated identification of potentially related terms for semantic mapping.

FGDC FEATURE REGISTRY MODEL (THESAURUS CONCEPT)

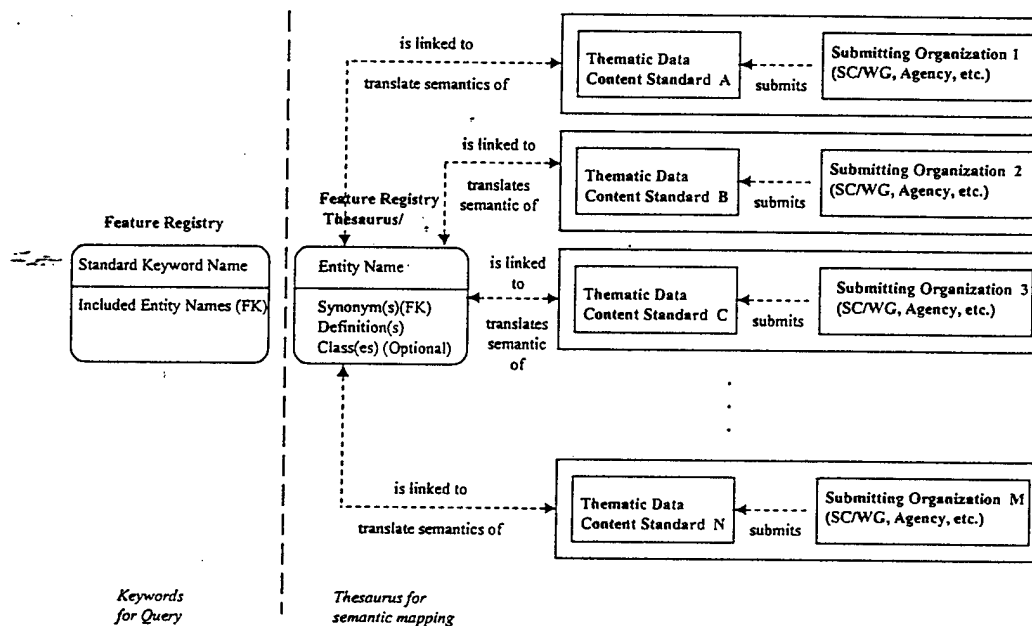Fig. 2. FGDC Feature Registry Data Model (Thesaurus Concept)

Applicable Research and Development

Semantic interoperability is a profound problem that is not limited to the geospatial community. The Internet has been described as a library without a card catalog. Search engines are becoming more sophisticated in information retrieval, but truly effective searches of relevant information are hampered by semantic differences between user communities. Research being performed includes how to map content or meaning, and the methods being researched are similar to the research efforts of the geospatial community.

The methods addressing semantic reconciliation fall into two broad categories: those of translation and those of standardization. Semantic reconciliation through translation seeks to provide a mapping between two disparate schemata. Semantic reconciliation through standardization seeks to provide a universal vocabulary.

The translation approach to semantic reconciliation involves the use of expert system technology. The system contains a catalog of the feature codes with their language descriptions for each coding system in the translation process. All available information is examined (extracting theme/table/feature/attribute information, or major/minor codes, for example) to construct a lineage for each feature. The rules base allows the mapping of the lineage information between schemata. Where ambiguity results, topology can be used for clarification. For example, it may be unclear whether a bridge feature is associated with a road or railroad. A proximity search of the topology would reveal the nature of the bridge.

The standardization approach to semantic reconciliation focuses on the creation of a library of terms across user communities. This approach is used by the FGDC's Feature Registry and Digital Library Initiative. It seeks to define a set of terms and meanings across communities. This approach is useful for searching federated databases for desired data. It also can be used to construct a catalog for a non-standard coding schema when using a translation approach to semantic reconciliation.

The difficulty associated with semantic interoperability of feature/attribute/domain information is exaggerated by the different languages and data structures used by geoprocessing systems. Complete

semantic interoperability would address these differences and solve the problem of application- and query-specific interoperability. Attempts have been made at defining general languages for geospatial processing that could provide for interoperability by allowing users to interact with many systems using a common, consistent language. And, converting data structures into a neutral format, generally of an object-oriented nature, is the basis of semantic translation. The object-oriented approach may provide the most promise. It facilitates the translation method and makes data structure differences meaningless. It is, perhaps, impracticable because of the prejudices of GIS manufacturers towards their chosen formats.

Need for Additional Research

Additional research is needed to support the automation of semantic mapping. There is need for research into what additional information should be added to feature catalogs (dictionaries), metadata, profiles, etc., to capture semantics that better express a community's meaning and intent. This research should lead to development of techniques, methods, procedures and/or guidance that support more automated mechanisms/methods for the discovery/determination of semantic mapping between communities' semantics.

Clearly an important step towards semantic interoperability has been research that discusses the need for information communities to document and agree upon their semantics (which also calls for consistency in how communities express their semantics.) When completed and available, the ISO TC211 Part 10, Feature Cataloging Methodology document will provide the necessary guidance for developing feature/attribute/domain catalogs. When implemented, this standard will help foster consistency between these feature dictionaries/catalog standards.

Another important step has been research that discusses the formalization of natural language definitions and expressions, as well as additional information that a community must capture regarding its semantics. However, this research must be distilled down to a clear set of directions that are understandable and useable by community experts as they document their semantics. In addition, this information should be documented as either an implementation or revision of ISO TC211, Part 10.

As mentioned previously, there also is a need for semantic mapping tools that can automate the identification of potentially related terms for semantic mapping, aided by the additional semantic information from an information community.

Conclusions

Semantic mapping is a very difficult process to automate. This is because the semantics are "wrapped up" in the intricacies of culture, natural languages, and human perception. The semantic mapping research area is akin to the area of image understanding and feature recognition (i.e., a human's ability to perform these tasks far exceed any computer's ability, however, automation is sought because these functions, as performed by humans, are very laborious and often yield inconsistent results). The authors believe that there are no complete solutions on the horizon that will solve either of these problems. Instead, both areas will be best addressed by dividing the problem into small, workable pieces, attacking the most significant pieces, and making small but meaningful steps towards semantic interoperability. The work being performed by the FGDC subcommittees and working groups to develop geospatial data standards that represent communities' semantics is a major step towards semantic interoperability within an information community. Publishing these standards and other significant feature catalogs, and addressing the relationship between the contents of these catalogs within the FGDC Feature Registry, will provide the roadmap for semantic mapping.

References

American National Standards Institute (ANSI). In National Committee for Information Technology Standards. Spatial Data Transfer Standard (ANSI NCITS 321:1999). New York, New York (1998)

Buehler, K., McKee, L. (ed.): The OpenGIS Guide: Introduction to Interoperable Geoprocessing, OpenGIS Consortium, Inc. Wayland, Massachusetts (1996)

Frank, A., Kuhn, W.: A Specification Language for Interoperable GIS. http://bbq.ncgia.ucsb.edu/conf/interop97/program/papers/frank.html

Gogolla, M., Hohenstien, U.: Towards a Semantic View of an Extended Entity Relationship Model. ACM Transactions of Database Systems. (1991) 16:369-416

Goodchild, M., Egenhofer, M., Fegeas, R.: Report of a Specialist Meeting Held Under the Auspices of the Varenius Project, Panel on Computational Implementation of Geographic Concepts. http://ncgia.uscb.edu/conf/interop97/interop_toc.html

International Organization for Standardization, Technical Committee for Geographic Information/Geomatics, Geographic Information – Part 10: Feature Cataloguing Methodology (ISO/Committee Draft 15046-10), Genève, Switzerland (1998)

Schatz, B., Chen, H.: Building Large-Scale Digital Libraries. http://computer.org/computer/dli/index.html

Sciore, E., Siegel, M., Rosenthal, A.: Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems. ACM Transactions on Database Systems. (1994) 19:254-90

Woodsford, P.: Data Conversion and Update in the Object Paradigm. http://www.lsl.co.uk/papers/datacon.htm

Yuan, M.: Development of a Global Conceptual Schema for Interoperable Geographic Information. http://bbq.ncgia.uscb.edu/con/interop97/program/papers/yuan/yuan.html